

Bayesian test of normality versus a Dirichlet process mixture alternative

Surya T. Tokdar

Department of Statistical Science

Duke University

tokdar@stat.duke.edu

Ryan Martin

Department of Mathematics, Statistics and Computer Science

University of Illinois at Chicago

rgmartin@math.uic.edu

August 16, 2011

Abstract

Testing if a p -dimensional sample, for $p \geq 1$, comes from a normal population is a fundamental problem in statistics. In this paper we propose a Bayesian test of p -variate normality against an alternative hypothesis characterized by a certain Dirichlet process mixture model. It is shown that this nonparametric alternative satisfies the desirable embedding and predictive matching properties with respect to the normal null model. To compute the Bayes factor, an efficient sequential importance sampler is proposed for evaluating the marginal likelihood under the nonparametric alternative. Numerical examples show that the proposed test has satisfactory discriminatory power when the distribution is not normal, and does not tend to over-fit when the distribution is normal.

Keywords and phrases: Bayes factor; embedding; goodness-of-fit test; importance sampling; nonparametric model; predictive matching.

1 Introduction

Consider a sample $X_{1:n} = (X_1, \dots, X_n)$ of n independent observations, $X_i \in \mathbb{R}^p$, distributed according to a common probability measure F , i.e., $F(A) = \Pr(X_i \in A)$ for measurable $A \subset \mathbb{R}^p$. Many statistical procedures, such as linear regression and analysis of variance, assume that the observations are samples from a normal population, so an important first step is verification of this assumption. The goal is to test the hypothesis $H_0 : F \in \mathcal{F}_0$, where

$$\mathcal{F}_0 = \{F_{\mu, \sigma} = \mathbf{N}(\mu, \sigma\sigma') : (\mu, \sigma) \in \mathbb{R}^p \times \mathbb{T}_p\} \quad (1)$$

is the set of all p -dimensional normal distributions. Here \mathbb{T}_p is the set of all $p \times p$ lower-triangular matrices with positive diagonal elements, so that $\sigma\sigma'$ is the Cholesky decomposition of the covariance matrix. A Bayesian treatment of this problem and, more generally, of goodness-of-fit testing of an arbitrary parametric model, has been recently explored in the literature (e.g., Berger and Guglielmi 2001; Carota and Parmigiani 1996; Florens et al. 1996; Verdinelli and Wasserman 1998). Unlike classical goodness-of-fit tests (cf. DasGupta 2008, Chap. 28), the Bayesian setup requires (i) a completion of the null model (1) with the assignment of a prior distribution Π_0 on \mathcal{F}_0 , determined by a possibly improper prior π_0 on $\mathbb{R}^p \times \mathbb{T}_p$ and the mapping $(\mu, \sigma) \mapsto F_{\mu, \sigma}$, and (ii) a specification of an alternative model $H_1 : F \in \mathcal{F}_1$ and a prior Π_1 on \mathcal{F}_1 . For a non-subjective treatment, it is natural not to restrict \mathcal{F}_1 to a finite-dimensional parametric family; instead, one chooses \mathcal{F}_1 to be an infinite-dimensional subset of probability measures on \mathbb{R}^p , and Π_1 a probability measure supported on \mathcal{F}_1 . Once the priors Π_0 and Π_1 are specified, one can report the Bayes factor

$$B(x_{1:n}) = \frac{\int_{\mathcal{F}_0} \left\{ \prod_{i=1}^n dF(x_i) \right\} d\Pi_0(F)}{\int_{\mathcal{F}_1} \left\{ \prod_{i=1}^n dF(x_i) \right\} d\Pi_1(F)} \quad (2)$$

as a measure of evidence against H_0 when $X_{1:n} = x_{1:n}$ are observed.

Towards a non-subjective, default choice of priors (π_0, Π_1) , both *embedding* and *predictive matching* have been proposed as desirable properties. Embedding typically means expressing Π_1 as a mixture $\int \Pi_{\mu, \sigma} d\pi_1(\mu, \sigma)$, where each $\Pi_{\mu, \sigma}$ gives a nonparametric prior distribution over probability measures on \mathbb{R}^p that form a local alternative to $F_{\mu, \sigma}$, and π_1 , like π_0 above, is a possibly improper prior on $\mathbb{R}^p \times \mathbb{T}_p$. The local alternative property of $\Pi_{\mu, \sigma}$ can be formalized by identifying $F_{\mu, \sigma}$ in (1) as the center of $\Pi_{\mu, \sigma}$, such as its mean (Berger and Guglielmi 2001; Carota and Parmigiani 1996), or some other measure of central tendency (Verdinelli and Wasserman 1998). Beyond this embedding property, it is difficult to pursue formal non-subjective requirements in choosing $\Pi_{\mu, \sigma}$ because of its extreme dimensionality. Instead, extrinsic justifications, such as computational ease and attractiveness of $\Pi_{\mu, \sigma}$ purely as a statistical model, are taken into consideration.

With embedding in place, and the choice of $\Pi_{\mu, \sigma}$ justified by extrinsic means, it remains to choose the priors π_0 and π_1 on $\mathbb{R}^p \times \mathbb{T}_p$. It is here that predictive matching plays an important role. Often a default choice of π_0 , usually improper, can be obtained through formal arguments, with inference on (μ, σ) under the parametric null model the ultimate goal (see Ghosh et al. 2006, Chapter 5). In light of the embedding property $\int F d\Pi_{\mu, \sigma}(F) = F_{\mu, \sigma}$, it is tempting to choose π_1 the same as π_0 (Carota and Parmigiani 1996) so that the elements of $\mathbb{R}^p \times \mathbb{T}_p$ are weighted the same under the null and alternative models. Berger and Guglielmi (2001) find this reasoning insufficient and argue that the choice $\pi_1 = \pi_0$ is partially justified if the predictive distribution of a hypothetical sample of size n_{\min} is the same under the two models, where n_{\min} is the minimal sample size needed to obtain a proper posterior for (μ, σ) under either model. The intuition is that a sample of size n_{\min} is needed to barely identify (μ, σ) , and one should not be able to tell the two models apart without additional data.

For univariate data $X_{1:n}$, Berger and Guglielmi (2001) present two compelling choices of $\Pi_{\mu, \sigma}$ in the form of Polya tree distributions (Lavine 1992, 1994; Mauldin et al. 1992) which satisfy the embedding property. For either choice, they demonstrate that, for the normal parametric model $N(\mu, \sigma^2)$, a common $\pi_0 = \pi_1 = \pi_H$, where π_H is the right Haar

measure given by $d\pi_H(\mu, \sigma) = (1/\sigma) d\mu d\sigma$ satisfies the predictive matching property with $n_{\min} = 2$. This is a non-trivial result that follows from a basic identity in Berger et al. (1998) and utilizes the fact that (μ, σ) can be identified as a location–scale parameter under both H_0 and H_1 , thus providing further extrinsic justification for choosing a common prior. The proposals in Berger and Guglielmi (2001) offer substantial improvements over a similar construction in Carota and Parmigiani (1996) based on a Dirichlet process distribution (Ferguson 1973) as $\Pi_{\mu, \sigma}$. That a Dirichlet process is supported on the set of discrete probability measures makes it an unattractive choice as an alternative to $N(\mu, \sigma^2)$ and leads to rather undesirable properties of the Bayes factor (Berger and Guglielmi 2001; Carota and Parmigiani 1996). Verdinelli and Wasserman (1998) use a logistic Gaussian process for $\Pi_{\mu, \sigma}$, which does support probability measures with smooth densities, but their proposal is difficult to compute with and poses serious challenges to a non-subjective treatment of the parametric priors π_0 and π_1 .

Interestingly, none of the available alternative specifications makes use of Dirichlet process mixtures, which are arguably the most attractive nonparametric distributions for modeling an unknown probability measure that admits a density; see, e.g., Müller and Quintana (2004) and the references therein. Models based on Dirichlet process mixture distributions, particularly those that mix over normal kernels, are easy to compute with, often via efficient Gibbs sampling methods (Escobar and West 1995; MacEachern and Müller 1998; MacEachern 1998; Neal 2000) and are known to possess optimal, adaptive, nearly parametric convergence rates in various applications, including density estimation (Ghosal et al. 1999, 2000; Ghosal and van der Vaart 2001, 2007). These nice convergence properties obtain because such a distribution sits on a space of probability measures with infinitely smooth densities, given by mixtures of normal probability measures, which offer sharp approximations to any probability measure with an arbitrarily smooth density. In contrast, a Polya tree distribution sits on probability measures with densities that are nowhere differentiable (Choudhuri et al. 2005), a property that automatically limits such a distribution’s ability to concentrate around probability measures with smooth densities and leads to inefficient estimation (Castillo 2008; van der Vaart and van Zanten 2008). Moreover, one can argue that for testing the fit of the normal distribution, probability measures with infinitely smooth densities form a more relevant alternative set than those with non-differentiable densities.

The main difficulty in using the Dirichlet process mixture distribution for $\Pi_{\mu, \sigma}$ appears to be the limited embedding capacity of such a distribution. For example, the mean of a Dirichlet process mixture of normals is a mixture of normals, and thus can equal $F_{\mu, \sigma}$ only if $F_{\mu, \sigma}$ is exactly a mixture of normals, not just a limit of such mixtures. This limited embedding capacity is the price one pays for ensuring that the nonparametric distribution concentrates on probability measures with smooth densities; logistic Gaussian processes suffer from the same shortcoming.

In this paper we show that for $F_{\mu, \sigma} = N(\mu, \sigma\sigma')$, a specially designed Dirichlet process mixture of normals distribution $\Pi_{\mu, \sigma}$ indeed satisfies the embedding property $\int F d\Pi_{\mu, \sigma}(F) = F_{\mu, \sigma}$. In addition to satisfying this technical condition and possessing the usual support properties of a Dirichlet process mixture of normals distribution, our special construction also offers an intuitive interpretation of $\Pi_{\mu, \sigma}$ as an alternative to $F_{\mu, \sigma}$. That is, a sample $F \sim \Pi_{\mu, \sigma}$ can be described as a random “granulation” of $F_{\mu, \sigma}$ into a mixture of normal measures, where each component occupies a fraction of the

volume of $F_{\mu,\sigma}$, with the volume being negatively correlated with the distance between the component's center and the center μ of $F_{\mu,\sigma}$. By introducing appropriate latent parameters, the mixture representation of F under the alternative translates into a latent clustering of the observations $X_{1:n}$ sampled from F (Escobar and West 1995), with the “extent” of clustering being a key factor in separating $F \sim \Pi_{\mu,\sigma}$ from the null element $F_{\mu,\sigma}$. By carefully choosing the model hyperparameters, we show how to perform the testing at different levels of separation between H_0 and H_1 by varying a single scalar parameter. We further show that, for $p \geq 1$, (μ, σ) remains a location–scale parameter for both $F_{\mu,\sigma}$ and $\Pi_{\mu,\sigma}$ and, consequently, a common choice of $\pi_0 = \pi_1 = \pi_H$, with $d\pi_H(\mu, \sigma) = (\det \sigma)^{-(p+1)/2} d\mu d\sigma$ giving the p -dimensional right Haar measure on $\mathbb{R}^p \times \mathbb{T}_p$, ensures that the predictive matching property holds with $n_{\min} = p + 1$ —one observation to identify μ and an additional p observations to identify σ . This last result, which is an extension of Theorem 4.1 in Berger et al. (1998), may be of independent interest.

The remainder of the paper is organized as follows. In Section 2 we introduce our particular class of Dirichlet process mixture model alternatives and derive the aforementioned properties. We develop a sequential importance sampling strategy in Section 3 for estimating the marginal likelihood under this Dirichlet process mixture model framework. Numerical examples are presented in Section 4, and concluding remarks are made in Section 5.

2 A Dirichlet process mixture alternative

2.1 Model specification

Let \mathbb{S}_p be the space of $p \times p$ symmetric positive definite matrices with all p eigenvalues in $(0, 1)$. For scalars ω_1 and ω_2 greater than $(p-1)/2$, let $\text{Be}(\omega_1, \omega_2)$ denote the multivariate beta distribution on \mathbb{S}_p (Muirhead 1982, Chapter 3.3) having density

$$\text{Be}(v \mid \omega_1, \omega_2) = a_p(\omega_1, \omega_2) (\det v)^{\omega_1 - (p+1)/2} \{\det(I_p - v)\}^{\omega_2 - (p+1)/2}, \quad (3)$$

where I_p is the $p \times p$ identity matrix and $a_p(\omega_1, \omega_2) = \Gamma_p(\omega_1 + \omega_2) / \Gamma_p(\omega_1) \Gamma_p(\omega_2)$, with $\Gamma_p(z) = \pi^{p(p-1)/4} \prod_{i=1}^p \Gamma(z - (i-1)/2)$ the p -variate gamma function. Write Ψ for the probability measure on $\mathbb{R}^p \times \mathbb{S}_p$ given by the law of (U, V) , where $V \sim \text{Be}(\omega_1, \omega_2)$ and $U \mid V \sim \text{N}(0, I_p - V)$. This law is well-defined, since $I_p - V \in \mathbb{S}_p$ with probability 1.

Let $\text{DP}(\alpha, \Psi)$ denote the Dirichlet process distribution with precision constant $\alpha > 0$ and base measure Ψ from above (Ferguson 1973). Recall that $\tilde{\Psi} \sim \text{DP}(\alpha, \Psi)$ means that for any positive integer k and any measurable partition B_1, \dots, B_k of $\mathbb{R}^p \times \mathbb{S}_p$, the probability vector $\{\tilde{\Psi}(B_1), \dots, \tilde{\Psi}(B_k)\}$ has a k -dimensional Dirichlet distribution with parameters $\{\alpha\Psi(B_1), \dots, \alpha\Psi(B_k)\}$. Given this Dirichlet process distribution, for each (μ, σ) , let $\text{DPM}_{\mu,\sigma}(\alpha, \Psi)$ denote the distribution of the random probability measure

$$\tilde{F}_{\mu,\sigma} = \int \text{N}(\mu + \sigma u, \sigma v \sigma') d\tilde{\Psi}(u, v), \quad \text{where } \tilde{\Psi} \sim \text{DP}(\alpha, \Psi); \quad (4)$$

see Lo (1984). This distribution is called a Dirichlet process mixture of normals, and it will be used in what follows as our local alternative $\Pi_{\mu,\sigma}$ for $F_{\mu,\sigma}$. The Dirichlet process

mixture (4) differs from that commonly found in the literature in its use of the normal–beta base measure Ψ as opposed to the more conventional normal–inverse Wishart base measure (Escobar and West 1995). The advantage is that this new formulation satisfies the desirable embedding property described in Section 1; see Theorem 1 below.

To summarize our model formulation, we have the following hierarchical specifications of the null and alternative models:

$$H_0 : X_{1:n} \mid (\mu, \sigma) \stackrel{\text{iid}}{\sim} F_{\mu, \sigma}, \quad (\mu, \sigma) \sim \pi_0 \quad (5)$$

$$H_1 : X_{1:n} \mid \tilde{F}_{\mu, \sigma} \stackrel{\text{iid}}{\sim} \tilde{F}_{\mu, \sigma}, \quad \tilde{F}_{\mu, \sigma} \mid (\mu, \sigma) \sim \text{DPM}_{\mu, \sigma}(\alpha, \Psi), \quad (\mu, \sigma) \sim \pi_1. \quad (6)$$

In the following subsections we shall show that the embedding and predictive matching properties hold for this formulation of the testing problem. We shall also give a clustering-based interpretation of H_1 as a natural alternative to H_0 , discuss the choice of beta hyperparameters (ω_1, ω_2) , and give some recommendations on Bayes factor reporting.

2.2 Null embedding property

Here we show that $\Pi_{\mu, \sigma} = \text{DPM}_{\mu, \sigma}(\alpha, \Psi)$, with (μ, σ) in $\mathbb{R}^p \times \mathbb{T}_p$, satisfies the embedding property, i.e., the mean of $\text{DPM}_{\mu, \sigma}(\alpha, \Psi)$ is $\mathbf{N}(\mu, \sigma\sigma')$. The key to the proof is that a normal distribution may be written as a convolution of two normals.

Theorem 1. *For any (μ, σ) , the mean of $\text{DPM}_{\mu, \sigma}(\alpha, \Psi)$ is $\mathbf{N}(\mu, \sigma\sigma')$.*

Proof. Given $\tilde{F}_{\mu, \sigma} \sim \text{DPM}_{\mu, \sigma}(\alpha, \Psi)$, let $\tilde{\Psi} \sim \text{DP}(\alpha, \Psi)$ be the corresponding random mixing distribution in (4). Since $\mathbf{E}(\tilde{\Psi}) = \Psi$, Fubini’s theorem implies

$$\begin{aligned} \mathbf{E}(\tilde{F}_{\mu, \sigma}) &= \int \mathbf{N}(\mu + \sigma u, \sigma v \sigma') d\mathbf{E}(\tilde{\Psi})(u, v) \\ &= \int \mathbf{N}(\mu + \sigma u, \sigma v \sigma') d\Psi(u, v) \\ &= \int \left\{ \int \mathbf{N}(\mu + \sigma u, \sigma v \sigma') d\mathbf{N}(u \mid I_d - v) \right\} d\text{Be}(v \mid \omega_1, \omega_2) \\ &= \int \mathbf{N}(\mu, \sigma\sigma') d\text{Be}(v \mid \omega_1, \omega_2) \\ &= \mathbf{N}(\mu, \sigma\sigma'). \end{aligned}$$

The next-to-last equality above follows from the well-known Gaussian convolution identity $\int \mathbf{N}(a + bu, s) d\mathbf{N}(u \mid c, t) = \mathbf{N}(a + bc, s + btb')$. \square

It is clear from the proof that the multivariate beta distribution is not absolutely necessary. In fact, the same proof goes through for any distribution for V supported on \mathbb{S}_p . So it may be possible to generalize our alternative formulation, though we will not investigate this any further here.

2.3 Predictive matching property

Theorem 1 establishes the crucial embedding property of the null within our proposed alternative. To show that predictive matching also holds in our formulation, we first establish a generalization of Theorem 4.1 of Berger et al. (1998) to any arbitrary dimension

$p \geq 1$ and a nonparametric prior on the underlying distribution F . Note that, for arbitrary p , the right Haar measure π_H on $\mathbb{R}^p \times \mathbb{T}_p$ is given by $d\pi_H(\mu, \sigma) = \prod_{j=1}^p \sigma_{jj}^{j-p-1} d\mu d\sigma$, where σ_{jj} is the j^{th} diagonal element of the matrix σ .

2.3.1 Predictive matching: a general result

In the following, for any random probability measure $F \sim \Pi$ on \mathbb{R}^p , let $m_{\Pi, n}$ denote the average product measure $\int F^{\times n} d\Pi(F)$, where $F^{\times n}$ denotes the n -fold product of F . Note that $m_{\Pi, n}$ is just the predictive distribution of a hypothetical sample of size n from the model $F \sim \Pi$ and $X_{1:n} \mid F \stackrel{\text{iid}}{\sim} F$. We shall call a family of probability measures $\{\Pi_{\mu, \sigma} : (\mu, \sigma) \in \mathbb{R}^p \times \mathbb{T}_p\}$ a location-scale family if there is a random probability measure F^* on \mathbb{R}^p such that, for any (μ, σ) , the random measure $F_{\mu, \sigma}^*$ defined as $dF_{\mu, \sigma}^*(x) = |\det \sigma|^{-1} dF^*(\sigma^{-1}(x - \mu))$ is distributed according to $\Pi_{\mu, \sigma}$. A location-scale family will be called rotation-invariant if, for any orthogonal matrix η , $F_{0, \eta}^*$ has the same distribution as F^* . Also, we shall call a location-scale family absolutely continuous if the characterizing F^* is absolutely continuous with respect to Lebesgue measure with probability 1.

Theorem 2. *Let $F \sim \Pi$ be a random probability measure on \mathbb{R}^p , with $\Pi = \int \Pi_{\mu, \sigma} d\pi_H(\mu, \sigma)$, where $\{\Pi_{\mu, \sigma} : (\mu, \sigma) \in \mathbb{R}^p \times \mathbb{T}_p\}$ is an absolutely continuous, rotation-invariant, location-scale family, and π_H is the right Haar measure on $\mathbb{R}^p \times \mathbb{T}_p$. Then, for any $x_{1:(p+1)}$ in \mathbb{R}^p such that $\tilde{x}_j = x_j - x_{p+1}$, $j \in 1 : p$, linearly independent,*

$$m_{\Pi, p+1}(x_{1:(p+1)}) = c_p^{-1} |\det \tilde{x}|^{-p}, \quad (7)$$

where \tilde{x} is the $p \times p$ matrix with columns $\tilde{x}_1, \dots, \tilde{x}_p$, and the normalizing constant $c_p = 2^p \pi^{p^2/2} / \Gamma_p(p/2)$ is the volume of the p -dimensional Steifel manifold.

In the proof below, integrals shall be carried out in the form of exterior products of differentials, which we denote as $(d\mu)$, etc. This use of exterior products leads to simpler change of variable formulas than those offered by traditional Jacobians. The changes of variable used below, and the corresponding exterior products, can be found in Muirhead (1982, Chap. 2).

Proof of Theorem 2. Let $F^* \sim \Pi^*$ be the random measure that characterizes the absolutely continuous, rotation-invariant, location-scale family $\Pi_{\mu, \sigma}$, and let f^* denote its Radon-Nikodym derivative with respect to Lebesgue measure on \mathbb{R}^p . By Fubini's theorem, $m_{\Pi, p+1}(x_{1:(p+1)})$ equals

$$\begin{aligned} & \int \left[\int_{\mathbb{R}^p \times \mathbb{T}_p} \left\{ \prod_{i=1}^{p+1} (\det \sigma)^{-1} f^*(\sigma^{-1}(x_i - \mu)) \right\} d\pi_H(\mu, \sigma) \right] d\Pi^*(f^*) \\ &= \int \left[\int_{\mathbb{R}^p \times \mathbb{T}_p} \left\{ \prod_{i=1}^{p+1} f^*(\sigma^{-1}(x_i - \mu)) \right\} (\det \sigma)^{-(p+1)} \prod_{i=1}^p \sigma_{ii}^{i-p+1} (d\mu)(d\sigma) \right] d\Pi^*(f^*). \end{aligned}$$

Write $I(f^*)$ for the integral over $\mathbb{R}^p \times \mathbb{T}_p$ inside the square brackets in the right-hand side of the above display. A change of variable $\tau = \sigma^{-1}$ implies τ ranges over \mathbb{T}_p , $\sigma_{ii} = \tau_{ii}^{-1}$, $\det \sigma = (\det \tau)^{-1}$, and $(d\sigma) = (\det \tau)^{-(p+1)} (d\tau)$. Therefore,

$$I(f^*) = \int_{\mathbb{R}^p \times \mathbb{T}_p} \left\{ \prod_{i=1}^{p+1} f^*(\tau(x_i - \mu)) \right\} \prod_{i=1}^p \tau_{ii}^{p+1-i} (d\mu)(d\tau).$$

By the rotation-invariant assumption, the right-hand side above remains unchanged if we replace F^\star with $F_{0,\eta'}^\star$, for any orthogonal matrix η . Let H , with $dH(\eta) = (\eta'd\eta)/c_p$, denote the Haar measure on \mathbb{O}_p , the space of $p \times p$ orthogonal matrices. Then we must have

$$I(f^\star) = c_p^{-1} \int_{\mathbb{R}^p \times \mathbb{T}_p \times \mathbb{O}_p} \left\{ \prod_{i=1}^{p+1} f^\star(\eta\tau(x_i - \mu)) \right\} \prod_{i=1}^p \tau_{ii}^{p+1-i}(d\mu)(d\sigma)(\eta'd\eta).$$

If we let $\nu = \eta\tau$, then ν ranges over the space \mathbb{G}_p of $p \times p$ non-singular matrices, $\det \tau = |\det \nu|$, and $(d\nu) = \prod_{i=1}^p \tau_{ii}^{p-i}(d\tau)(\eta'd\eta)$. Therefore,

$$I(f^\star) = c_p^{-1} \int_{\mathbb{R}^p \times \mathbb{G}_p} \left\{ \prod_{i=1}^{p+1} f^\star(\nu(x_i - \mu)) \right\} |\det \nu| (d\mu)(d\nu).$$

Note that (μ, ν) effectively ranges over $\mathbb{R}^{p \times (p+1)}$, the $(p+1)$ -fold product of \mathbb{R}^p . Make a final change of variable, $z_i = \nu(x_i - \mu)$, $i \in 1 : (p+1)$. The inverse transformation is given by $\nu = \tilde{z}\tilde{x}^{-1}$, $\mu = x_{p+1} - \tilde{x}\tilde{z}^{-1}z_{p+1}$, where \tilde{x} is as in the statement of the theorem and, likewise, \tilde{z} is the $p \times p$ matrix with columns $\tilde{z}_i = z_i - z_{p+1}$. Therefore, the Jacobian of this transformation equals $|\det \tilde{z}| |\det \tilde{x}|^{-(p-1)}$ and so

$$I(f^\star) = c_p^{-1} \int_{\mathbb{R}^{p \times (p+1)}} \left\{ \prod_{i=1}^{p+1} f^\star(z_i) \right\} |\det \tilde{x}|^{-p} dz_{1:(p+1)} = c_p^{-1} |\det \tilde{x}|^{-p},$$

since $\int f^\star(z_i) dz_i = 1$ with Π^\star -probability 1 for each $i \in 1 : (p+1)$. The claim (7) now follows immediately since $I(f^\star)$ is free of f^\star . \square

In particular, in the univariate case $p = 1$, the minimum training sample size is $n_{\min} = 2$ and such a training sample consists of two distinct observations, say, x_1 and x_2 . Then \tilde{x} is just a number, namely $x_1 - x_2$, and $|\det \tilde{x}| = |x_1 - x_2|$. Furthermore, it is easy to check from the formula that $c_1 = 2$. Therefore, the predictive density for $x_{1:2}$ is simply $\{2|x_1 - x_2|\}^{-1}$ which is exactly the result given in Berger et al. (1998, page 309, equation 7). Their Theorem 4.1 extends this result to the linear model case. An extension of our Theorem 2 to the linear model case is surely possible, but we do not pursue this direction any further here.

The important point is that, in addition to an extension to $p > 1$ dimensions, Theorem 2 allows for a nonparametric prior on the underlying distribution F . The condition on this nonparametric prior is that, given (μ, σ) , a draw $\tilde{F}_{\mu,\sigma} \sim \Pi_{\mu,\sigma}$ will almost surely admit an absolutely continuous, rotation-invariant, location-scale representation. Berger and Guglielmi (2001) argue that, for the $p = 1$ case, their Polya tree prior satisfies this property. In the next subsection we show that the proposed family $\{\text{DPM}_{\mu,\sigma}(\alpha, \Psi) : (\mu, \sigma) \in \mathbb{R}^p \times \mathbb{T}_p\}$ satisfies the conditions of Theorem 2 for general p .

2.3.2 Predictive matching for the Dirichlet mixture alternative

Next we apply Theorem 2 to the particular Dirichlet process mixture model of Section 2.1 to establish the advertised predictive matching property. In particular, it follows immediately from the definition (4) that $\text{DPM}_{\mu,\sigma}(\alpha, \Psi)$ is a location-scale family characterized

by the random measure $F^\star = \int \mathbf{N}(u, v) d\tilde{\Psi}(u, v)$ with $\tilde{\Psi} \sim \text{DP}(\alpha, \Psi)$. Clearly F^\star is absolutely continuous with respect to Lebesgue measure because each $\mathbf{N}(u, v)$ is so. The following lemma shows that this family is also rotation-invariant.

Lemma 1. *For $F^\star = \int \mathbf{N}(u, v) d\tilde{\Psi}(u, v)$ with $\tilde{\Psi} \sim \text{DP}(\alpha, \Psi)$ and any $\eta \in \mathbb{O}_p$, both F^\star and $F_{0, \eta'}^\star$ have the same distribution.*

Proof. For $\eta \in \mathbb{O}_p$, $d\mathbf{N}(\eta x \mid u, v) = d\mathbf{N}(x \mid \eta' u, \eta' v \eta)$ and, therefore,

$$F_{0, \eta'}^\star = \int \mathbf{N}(u, v) d\tilde{\Psi}_\eta(u, v), \quad \tilde{\Psi}_\eta \sim \text{DP}(\alpha, \Psi_\eta),$$

where Ψ_η denotes the law of $(U_\eta, V_\eta) = (\eta' U, \eta' V \eta)$ when $(U, V) \sim \Psi$. But if $V \sim \text{Be}(\omega_1, \omega_2)$, then also $V_\eta \sim \text{Be}(\omega_1, \omega_2)$ (see Muirhead 1982, Exercise 3.22d) and if $U \mid V \sim \mathbf{N}(0, I_p - V)$, then $U_\eta \mid V_\eta \sim \mathbf{N}(0, I_p - V_\eta)$. Therefore, by construction of Ψ , we have $\Psi_\eta = \Psi$ and, hence, F^\star and $F_{0, \eta'}^\star$ have the same distribution. \square

It is straightforward to see that $\{\langle F_{\mu, \sigma} \rangle : (\mu, \sigma) \in \mathbb{R}^p \times \mathbb{T}_p\}$, where $\langle F \rangle$ denotes a degenerate distribution at F , is also an absolutely continuous, rotation-invariant, location-scale family. This leads to the following predictive matching property.

Theorem 3. *The two models (5) and (6), with $\pi_0 = \pi_1 = \pi_H$, produce the same predictive distribution for any hypothetical sample of size $n_{\min} = p + 1$. That is, models (5) and (6) satisfy the predictive matching property.*

Proof. The claim follows from Theorem 2 since both $\{\text{DPM}_{\mu, \sigma}(\alpha, \Psi) : (\mu, \sigma) \in \mathbb{R}^p \times \mathbb{T}_p\}$ and $\{\langle F_{\mu, \sigma} \rangle : (\mu, \sigma) \in \mathbb{R}^p \times \mathbb{T}_p\}$ are absolutely continuous, rotation-invariant, location-scale families, and the set of all $(x_1, \dots, x_{p+1}) \in \mathbb{R}^{p \times (p+1)}$ with singular \tilde{x} matrix forms a null set with respect to Lebesgue measure. \square

2.4 Characterization via latent clusters

The stick-breaking representation of a Dirichlet process (Sethuraman 1994) states that a random $\tilde{\Psi} \sim \text{DP}(\alpha, \Psi)$ can be written as

$$\tilde{\Psi} = \sum_{h=1}^{\infty} q_h \langle (U_h, V_h) \rangle, \quad (8)$$

where (U_h, V_h) , $h \geq 1$, are independently draws from Ψ , $q_h = \beta_h \prod_{j < h} (1 - \beta_j)$ where β_h , $h \geq 1$, are independent draws from a univariate $\text{Be}(1, \alpha)$ distribution, and again $\langle (U, V) \rangle$ denotes a degenerate distribution at (U, V) . The vector $q_{1:\infty}$ forms a probability vector, i.e., $q_h \geq 0$ and $\sum_h q_h = 1$, with probability 1.

Consequently, given (μ, σ) , a draw from the local Dirichlet process mixture alternative $\text{DPM}_{\mu, \sigma}(\alpha, \Psi)$ has a representation of the form

$$\tilde{F}_{\mu, \sigma} = \sum_{h=1}^{\infty} q_h \mathbf{N}(\mu + \sigma U_h, \sigma V_h \sigma'), \quad (9)$$

with (q_h, U_h, V_h) , $h \geq 1$, as described above. Therefore, given (μ, σ) , the local alternative $X_{1:n} \stackrel{\text{iid}}{\sim} \tilde{F}_{\mu, \sigma}$ is equivalent to saying that the X_i 's are independently distributed

according to $\mathbf{N}(\mu + \sigma U_{h_i}, \sigma V_{h_i} \sigma')$ where the h_i 's are randomly drawn labels with $\Pr(h_i = h) = q_h$. Because the h_i 's can have ties with positive probability, the equivalence relation $i \sim j$ if and only if $h_i = h_j$ partitions the data $X_{1:n}$ into clusters of observations (Ghosh and Ramamoorthi 2003, Chap. 3), where the X_i 's in a cluster are independent $\mathbf{N}(\mu + \sigma U, \sigma V \sigma')$ observations, with $(U, V) \sim \Psi$. The center of this cluster is at a σU shift from the center μ of the corresponding null element $\mathbf{N}(\mu, \sigma \sigma')$ and occupies a $(\det V)^{1/2} \in (0, 1)$ fraction of the volume of the null element. The magnitude $(U'U)^{1/2}$ of the shift (relative to σ) is stochastically inversely related to the volume fraction $(\det V)^{1/2}$, as can be seen in the following theorem.

Theorem 4. *If $(U, V) \sim \Psi$, then $\text{Cov}(U'U, \det V) \leq 0$.*

Proof. Since $U|V \sim \mathbf{N}(0, I_p - V)$, it follows that $\mathbf{E}(U'U | V) = \text{tr } \mathbf{E}(UU' | V) = \text{tr}(I_p - V) = p - \text{tr } V$, where $\text{tr } A$ returns the trace of a symmetric matrix A . Then

$$\text{Cov}(U'U, \det V) = \text{Cov}\{\mathbf{E}(U'U | V), \det V\} = -\text{Cov}(\text{tr } V, \det V). \quad (10)$$

According to Muirhead (1982, p. 112), the eigenvalues of $V \sim \text{Be}(\omega_1, \omega_2)$ are distributionally equivalent to the eigenvalues of $A(A + B)^{-1}$, where $A \sim \text{Wish}(\omega_1, I_p)$ and $B \sim \text{Wish}(\omega_2, I_p)$. Since $\text{tr } V$ and $\det V$ are both coordinate-wise increasing functions of these eigenvalues, it follows from the main result of Dykstra and Hewett (1978, Sec. 5) that $\text{Cov}(\text{tr } V, \det V) \geq 0$. This, along with (10), completes the proof. \square

Therefore, given (μ, σ) , the local alternative $\tilde{F}_{\mu, \sigma}$ in (9) can be seen as local granulations of a population of fine particles evenly distributed according to $\mathbf{N}(\mu, \sigma \sigma')$. The local granulations form clusters with bell-shaped curves, each occupying only a fraction of the total volume of the population. The further the cluster center is from the original $\mathbf{N}(\mu, \sigma \sigma')$ population center, the smaller the cluster size is likely to be.

2.5 Choice of beta hyperparameters

It is well-known that, in (8), the degree of clustering, i.e., the prevalence of ties in the labels h_i , is controlled by the precision parameter α ; see, e.g., Ghosh and Ramamoorthi (2003, Chap. 3). It follows from our discussion above that the degree of clustering is a key determinant of how different a local alternative $F \sim \text{DPM}_{\mu, \sigma}(\alpha, \Psi)$ is from $F_{\mu, \sigma}$. For this reason, we choose to use α as a tuning parameter that encodes the separation between $\mathbf{N}(\mu, \sigma \sigma')$ and a draw from $\text{DPM}_{\mu, \sigma}(\alpha, \Psi)$. By varying α , we aim to cover a large spectrum of separation of the Dirichlet process mixture alternative and the null model. Operationally, the testing will be performed at every α in a range $(0, \infty)$ and the whole range of Bayes factors will be reported, from which the user can choose any summary of evidence against the null, such as those obtained by maximizing or averaging over α ; see Berger and Guglielmi (2001) for more discussion on the use of such tuning parameters in testing.

For large values of α , each cluster weight q_h becomes miniscule. This discourages ties among the latent labels and makes $\tilde{\Psi}$ a fine discrete approximation of the continuous measure Ψ . In fact, as $\alpha \rightarrow \infty$, the random measure $\tilde{\Psi}$ collapses to the fixed measure Ψ . Consequently, given (μ, σ) , the random measure $\tilde{F}_{\mu, \sigma}$ in (9) converges to $F_{\mu, \sigma} = \mathbf{N}(\mu, \sigma \sigma')$. On the other hand, as $\alpha \rightarrow 0$, the random $\tilde{\Psi}$ converges to a random degenerate

distribution $\langle(U, V)\rangle$ and, hence, $\tilde{F}_{\mu,\sigma}$ converges to $\mathbf{N}(\mu + \sigma U, \sigma V \sigma')$, with $(U, V) \sim \Psi$. Thus, elements of the null model appear as the limit of $\tilde{F}_{\mu,\sigma}$ when one turns the α knob to zero. The random nature of this limit, however, is unappealing because $\tilde{F}_{\mu,\sigma}$ is specifically designed to provide an alternative to $F_{\mu,\sigma}$ only. This shortcoming goes away if we ensure that Ψ also converges to $\langle(0, I_p)\rangle$ in the limit, so that $\tilde{F}_{\mu,\sigma}$ converges to its null counterpart $F_{\mu,\sigma}$. For Ψ to converge to $\langle(0, I_p)\rangle$, one needs $\omega_1/(\omega_1 + \omega_2) \rightarrow 1$ as $\alpha \rightarrow 0$. To ensure this, we choose

$$\omega_1 = \frac{p+1}{2} + \alpha^{-(p+1)/2} \quad \text{and} \quad \omega_2 = \frac{p+1}{2} + \alpha^{(p+1)/2}. \quad (11)$$

This is only one of the many possible ways to accomplish $\omega_1/(\omega_1 + \omega_2) \rightarrow 1$ as $\alpha \rightarrow 0$, which is essential to make Ψ collapse to $\langle(0, I_p)\rangle$. In addition to these theoretical justifications, the particular choice in (11) is simple and useful from a computational point of view.

Note also that, for (ω_1, ω_2) in (11), $\omega_1/(\omega_1 + \omega_2) \rightarrow 0$ as $\alpha \rightarrow \infty$. Consequently, turning the α knob to ∞ also results in the null model in the limit. However, the nature of this limiting path is quite different from that when $\alpha \rightarrow 0$. Indeed, for small non-zero α , the random $\tilde{\Psi}$ is close to degenerate. In terms of the stick-breaking representation (8), a single q_h dominates the rest. For large but finite α , there are no dominating q_h 's, and $\tilde{F}_{\mu,\sigma}$ is made up of small contributions from many normal components. By choosing ω_1 and ω_2 as in (11), the variance parameter $\sigma V_h \sigma'$ of any such component is made increasingly tiny as α increases. Consequently, $\tilde{F}_{\mu,\sigma}$ is an approximately continuous mixture of many narrow normal kernels with an overall shape resembling $\mathbf{N}(\mu, \sigma \sigma')$.

2.6 The precision parameter and Bayes factor reporting

With the complete formulation of the null and alternative models, we can rewrite the Bayes factor in (2) as

$$B = \frac{\int_{\mathbb{R}^p \times \mathbb{T}_p} \left\{ \prod_{i=1}^n dF_{\mu,\sigma}(x_i) \right\} d\pi_H(\mu, \sigma)}{\int_{\mathbb{R}^p \times \mathbb{T}_p} \left\{ \prod_{i=1}^n dF(x_i) \right\} d\text{DPM}_{\mu,\sigma}(F \mid \alpha, \Psi) d\pi_H(\mu, \sigma)}. \quad (12)$$

One can either subject the Bayes factor itself to a threshold, rejecting H_0 if and only if B is too small, as in Jeffreys (1961, page 432), or do the same with the posterior odds rB , where r is one's postulated prior odds in favor of H_0 .

The Bayes factor above actually depends on α and (ω_1, ω_2) . With (ω_1, ω_2) chosen as in (11), the Dirichlet process mixture alternative, and hence the Bayes factor, is entirely determined by the scalar α . As in Berger and Guglielmi (2001), we recommend carrying out the goodness-of-fit test separately for a range of α values covering its range, and presenting the Bayes factors side by side in the form of a plot. In our examples, we consider a range of α values comparable to the interval (n^{-1}, n^2) suggested by Escobar (1994). This plot of B versus α can also be interpreted as the reciprocal marginal likelihood for α under the Dirichlet process mixture model formulation. The minimum of this plot, corresponding to the Type II maximum marginal likelihood estimate of α , gives the maximum possible evidence against the null within the given class of alternatives. Alternatively, it is possible to combine these α -indexed family of models into a single overarching model by incorporating a prior distribution (e.g., gamma) for α .

3 Computation

The computation of the numerator of $B(x_{1:n})$ in (12) is simple since the posterior under the null has a nice form. Computation of the denominator, on the other hand, requires integration over an infinite-dimensional space and is non-trivial; see Kass and Raftery (1995). We compute the denominator by introducing latent parameter values $(U, V)_{1:n}$ and applying a variation on the sequential imputation technique of Liu (1996). Our method differs from Liu's in that (i) we partially collapse the mixture model in its U component, and (ii) we deal with the outer integration over (μ, σ) .

The behavior of $X_{1:n}$ under our mixture specification of H_1 can be described as

$$X_{1:n} \mid \{S_{1:n}, (U^*, V^*)_{1:n}, \mu, \sigma\} \sim \prod_{i=1}^n \mathbf{N}(X_i \mid \mu + \sigma U_{S_i}^*, \sigma V_{S_i}^* \sigma'), \quad (13)$$

where $(\mu, \sigma) \sim \pi_H(\mu, \sigma)$, $(U^*, V^*)_{1:n} \stackrel{\text{iid}}{\sim} \Psi$ are latent mixing parameters and $S_{1:n}$ is a latent vector of labels that tracks which observation is allocated to which mixing component, and these three variables are mutually independent. It suffices to restrict the latent labels to the space $\{(s_1, \dots, s_n) \in (1:n)^n : s_1 = 1, s_{i+1} \leq \max(s_{1:i}) + 1, i \in 1:(n-1)\}$. From the Polya urn representation (Blackwell and MacQueen 1973) of a Dirichlet process, the distribution of $S_{1:n}$ can be written as

$$\Pr(S_1 = 1) = 1, \quad \Pr(S_{i+1} = \ell \mid S_{1:i}) = \begin{cases} \frac{k_\ell(S_{1:i})}{\alpha + i} & \ell \in 1 : \max(S_{1:i}) \\ \frac{\alpha}{\alpha + i} & \ell = \max(S_{1:i}) + 1. \end{cases}$$

where $k_\ell(S_{1:i}) = |S_{1:i} = \ell|$ counts the number of $j \in 1:i$ with $S_j = \ell$.

It is possible to integrate out $U_{1:n}^*$ from this description, with suitable changes made to (13). Let $f(x_{1:n}, s_{1:n}, v_{1:n}^*, \mu, \sigma)$ denote the resulting joint density of $(X_{1:n}, S_{1:n}, V_{1:n}^*, \mu, \sigma)$ and, for $i \in 0:(n-1)$, let $f_{i+1}^X(x_{i+1} \mid x_{1:i}, s_{1:i}, v_{1:n}^*, \mu, \sigma)$ denote the associated conditional density of X_{i+1} given $(X_{1:i}, S_{1:i}, V_{1:n}^*, \mu, \sigma)$. Also let $f_{i+1}^S(s_{i+1} \mid x_{1:i+1}, s_{1:i}, v_{1:n}^*, \mu, \sigma)$ denote the conditional density of S_{i+1} given $(X_{1:i+1}, S_{1:i}, V_{1:n}^*, \mu, \sigma)$. It is easy to check that

$$f_{i+1}^X(x_{i+1} \mid x_{1:i}, s_{1:i}, v_{1:n}^*, \mu, \sigma) = \frac{\alpha}{\alpha + i} \mathbf{N}(x_{i+1} \mid \mu, \sigma \sigma') + \sum_{\ell=1}^{\max(s_{1:i})} \frac{k_\ell(s_{1:i})}{\alpha + i} \mathbf{N}(x_{i+1} \mid \mu_\ell, \sigma_\ell \sigma'_\ell) \quad (14)$$

$$f_{i+1}^S(\ell \mid x_{1:i+1}, s_{1:i}, v_{1:n}^*, \mu, \sigma) = \begin{cases} c^{-1} k_\ell(s_{1:i}) \mathbf{N}(x_{i+1} \mid \mu_\ell, \sigma_\ell \sigma'_\ell), & \ell \in 1 : \max(s_{1:i}) \\ c^{-1} \alpha \mathbf{N}(x_{i+1} \mid \mu, \sigma \sigma'), & \ell = \max(s_{1:i}) + 1 \end{cases} \quad (15)$$

with

$$\begin{aligned} \mu_\ell &= \mu + \sigma(I_p - v_\ell^*) \{v_\ell^* + k_\ell(s_{1:i})(I_p - v_\ell^*)\}^{-1} \sum_{j=1}^i x_j 1(s_j = \ell), \\ \sigma_\ell \sigma'_\ell &= \sigma v_\ell^* \{v_\ell^* + k_\ell(s_{1:i})(I_p - v_\ell^*)\}^{-1} \{I_p + k_\ell(s_{1:i})(I_p - v_\ell^*)\} \sigma', \end{aligned} \quad (16)$$

and $c = \alpha \mathbf{N}(x_{i+1} \mid \mu, \sigma \sigma') + \sum_{\ell=1}^{\max(s_{1:i})} k_\ell(s_{1:i}) \mathbf{N}(x_{i+1} \mid \mu_\ell, \sigma_\ell \sigma'_\ell)$.

The marginal density $f_{H_1}(x_{1:n})$ of $X_{1:n}$ under H_1 can be calculated by integrating $f(x_{1:n}, s_{1:n}, v_{1:n}^*, \mu, \sigma)$ with respect to $(s_{1:n}, v_{1:n}^*, \mu, \sigma)$. While this integral is analytically

intractable, one can approximate it by importance sampling Monte Carlo (Liu 2001, Chap. 2.5). Let $(S_{1:n}^m, V_{1:n}^{\star m}, \mu^m, \sigma^m)$, $m \in 1 : M$, be independent draws from a joint density $f_{\text{imp}}(s_{1:n}, v_{1:n}^{\star}, \mu, \sigma)$ on the space of $(S_{1:n}, V_{1:n}^{\star}, \mu, \sigma)$. Then a root- M consistent estimate of the marginal likelihood $f_{H_1}(x_{1:n})$ is

$$\hat{f}_{H_1}(x_{1:n}) = \frac{1}{M} \sum_{m=1}^M \frac{f(x_{1:n}, S_{1:n}^m, V_{1:n}^{\star m}, \mu^m, \sigma^m)}{f_{\text{imp}}(S_{1:n}^m, V_{1:n}^{\star m}, \mu^m, \sigma^m)}. \quad (17)$$

The efficiency of this approximation depends on how well $f_{\text{imp}}(s_{1:n}, v_{1:n}^{\star}, \mu, \sigma)$ approximates the conditional density of $(S_{1:n}, V_{1:n}^{\star}, \mu, \sigma)$, given $X_{1:n} = x_{1:n}$, under the joint density $f(x_{1:n}, s_{1:n}, v_{1:n}^{\star}, \mu, \sigma)$. Below we present an f_{imp} that achieves a fairly good approximation.

Let $f_{\text{imp}}(s_{1:n}, v_{1:n}^{\star}, \mu, \sigma)$ be the joint density of $(S_{1:n}, V_{1:n}^{\star}, \mu, \sigma)$ where (μ, σ) has density $f_{\text{imp}}^{\mu, \sigma}(\mu, \sigma)$ to be specified later, $V_{1:n}^{\star} \stackrel{\text{iid}}{\sim} \text{Be}(\omega_1, \omega_2)$ independently of (μ, σ) , and $S_{1:n}$ given $V_{1:n}^{\star} = v_{1:n}^{\star}$ and (μ, σ) has density $\prod_{i=0}^{n-1} f_{i+1}^S(s_{i+1} \mid x_{1:i+1}, s_{1:i}, v_{1:n}^{\star}, \mu, \sigma)$ as given in (15). This choice can be justified on two accounts. First, the conditional importance density of S_{i+1} given $(S_{1:i}, V_{1:n}^{\star}, \mu, \sigma)$ is precisely the partial conditional density of S_{i+1} given $(X_{1:(i+1)}, S_{1:i}, V_{1:n}^{\star}, \mu, \sigma)$ under f . Second, the partial conditional density under f of V_{ℓ}^{\star} given $\{S_{i+1} = \max(S_{1:i}) + 1 = \ell, X_{1:i+1}, V_{1:l-1}^{\star}, \mu, \sigma\}$ is precisely $\text{Be}(\omega_1, \omega_2)$. Let f^V and $f^{\mu, \sigma}$ denote the densities of $V_{1:n}^{\star} \stackrel{\text{iid}}{\sim} \text{Be}(\omega_1, \omega_2)$ and $(\mu, \sigma) \sim \pi_H$. It can be shown that

$$\begin{aligned} f(x_{1:n}, s_{1:n}, v_{1:n}^{\star}, \mu, \sigma) &= f^{\mu, \sigma}(\mu, \sigma) f^V(v_{1:n}^{\star}) \\ &\quad \times \prod_{i=0}^{n-1} \{f_{i+1}^X(x_{i+1} \mid x_{1:i}, s_{1:i}, v_{1:n}^{\star}, \mu, \sigma) f_{i+1}^S(s_{i+1} \mid x_{1:i+1}, s_{1:i}, v_{1:n}^{\star}, \mu, \sigma)\} \\ &= f_{\text{imp}}(s_{1:n}, v_{1:n}^{\star}, \mu, \sigma) \frac{f^{\mu, \sigma}(\mu, \sigma)}{f_{\text{imp}}^{\mu, \sigma}(\mu, \sigma)} \prod_{i=0}^{n-1} f_{i+1}^X(x_{i+1} \mid x_{1:i}, s_{1:i}, v_{1:n}^{\star}, \mu, \sigma) \end{aligned}$$

where the first equality follows from the sequential imputation calculations of Liu (1996) and the second equality follows from the definition of f_{imp} . Therefore (17) simplifies to

$$\begin{aligned} \hat{f}_{H_1}(x_{1:n}) &= \frac{1}{M} \sum_{m=1}^M \frac{f^{\mu, \sigma}(\mu^m, \sigma^m)}{f_{\text{imp}}^{\mu, \sigma}(\mu^m, \sigma^m)} \prod_{i=0}^{n-1} f_{i+1}^X(x_{i+1} \mid x_{1:i}, s_{1:i}^m, v_{1:n}^{\star}, \mu, \sigma) \\ &= \frac{1}{M} \sum_{m=1}^M \frac{f_{H_0}(x_{1:n}) f_{H_0}^{\mu, \sigma}(\mu^m, \sigma^m \mid x_{1:n})}{f_{\text{imp}}(\mu^m, \sigma^m)} \prod_{i=0}^{n-1} \frac{f_{i+1}^X(x_{i+1} \mid x_{1:i}, s_{1:i}^m, v_{1:n}^{\star}, \mu, \sigma)}{\mathbf{N}(x_{i+1} \mid \mu^m, \sigma^m \sigma'^m)} \end{aligned} \quad (18)$$

where $f_{H_0}(x_{1:n}) = \int \prod_{i=1}^n \mathbf{N}(x_i \mid \mu, \sigma \sigma') d\pi_H(\mu, \sigma)$ denotes the marginal density of $X_{1:n}$ under H_0 and $f_{H_0}^{\mu, \sigma}(\mu, \sigma \mid x_{1:n})$ denotes the posterior density of (μ, σ) under this model.

With $f_{H_1}(x_{1:n})$ estimated by $\hat{f}_{H_1}(x_{1:n})$ in (18), an estimate of B in (12) obtains in $\hat{B} = f_{H_0}(x_{1:n}) / \hat{f}_{H_1}(x_{1:n})$. Due to (14) and (18), this estimate simplifies to

$$\hat{B} = \left[\frac{1}{M} \sum_{m=1}^M \frac{f_{H_0}^{\mu, \sigma}(\mu^m, \sigma^m \mid x_{1:n})}{f_{\text{imp}}(\mu^m, \sigma^m)} \prod_{i=0}^{n-1} \left\{ \frac{\alpha}{\alpha + i} + \sum_{\ell=1}^{\max(s_{1:i}^m)} \frac{k_{\ell}(s_{1:i}^m) \mathbf{N}(x_{i+1} \mid \mu_{\ell}^m, \sigma_{\ell}^m \sigma'_{\ell}^m)}{\alpha + i \mathbf{N}(x_{i+1} \mid \mu^m, \sigma^m \sigma'^m)} \right\} \right]^{-1} \quad (19)$$

with formulas for μ_ℓ^m and σ_ℓ^m suitably adapted from (16). In our implementations, we use the approximation in (19), where for every m , we process the observations $x_{1:n}$ in a random order. This extra randomness does not violate the theoretical validity of the approximation, instead, makes it practically more efficient.

Equation (19) may suggest that $f_{\text{imp}}^{\mu,\sigma}(\mu, \sigma)$ can be chosen to approximate $f_{H_0}^{\mu,\sigma}(\mu, \sigma \mid x_{1:n})$. In reality, it should be chosen to approximate $f_{H_1}^{\mu,\sigma}(\mu, \sigma \mid x_{1:n})$ the posterior density of (μ, σ) under H_1 to make (17) an efficient approximation. However, due to the embedding and predictive matching properties of the alternative, the posterior densities of (μ, σ) under the two models can be expected to be similar to each other. Therefore a reasonable choice of $f_{\text{imp}}^{\mu,\sigma}(\mu, \sigma)$ seems an approximation to $f_{H_0}^{\mu,\sigma}(\mu, \sigma \mid x_{1:n})$ with heavier tails to guard against possible mismatch between this density and $f_{H_1}^{\mu,\sigma}(\mu, \sigma \mid x_{1:n})$. In the $p = 1$ case, there are a number of standard ways this can be done. First, like in Berger and Guglielmi (2001), one can use the sampling distribution of the maximum likelihood estimates $(\hat{\mu}, \hat{\sigma})$ to produce a bivariate Student-t density for $(\mu, \log \sigma)$ from which importance samples can be easily obtained. In the examples presented in Section 4.1, we take a slightly simpler approach. Specifically, $f_{\text{imp}}^{\mu,\sigma}$ is the density of (μ, σ) where, given σ , $n^{-1/2}(\mu - \hat{\mu})/\sigma$ has a Student-t distribution with 3 degrees of freedom, and σ has a Burr distribution with density function $(1 + \sigma/\hat{\sigma})^{-2}$. The $p > 1$ case requires samples of mean vectors and covariance matrices as opposed to scalars. For this we take $f_{\text{imp}}^{\mu,\sigma}$ to be the multivariate normal-inverse Wishart posterior density under H_0 , but rescaled to have heavier tails. In particular, we scale the covariance matrix of the normal component by a factor of n , and take the degrees of freedom of the inverse Wishart component to be $\max\{p, n - pn^{1/2}\}$. This approach is suitable for our general purposes, but further fine-tuning would likely lead to improved efficiency. R code is available at the first author's website, www.stat.duke.edu/~st118/Software.

4 Numerical illustrations

4.1 Univariate case, $p = 1$

Example 1. As a first illustration we revisit the example presented in Berger and Guglielmi (2001) with observations $X_{1:n}$ on the log-lifetimes of $n = 100$ Kevlar pressure vessels (Andrews and Herzberg 1985, page 183). A histogram in Figure 1(a) reveals that this data set has a significant left-skew, so we are inclined to believe that the underlying distribution is non-Gaussian. But we shall evaluate the Bayes factor using the sequential importance sampling strategy in Section 3 to confirm our inclination. In particular, for 13 choices of α , namely $\alpha = 2^a$, $a \in -6 : 6$, the marginal likelihood approximation (17) is obtained based on a Monte Carlo sample size of $M = 20,000$. Figure 1 shows a plot of the corresponding Bayes factor, on the \log_{10} scale, as a function of $a = \log_2(\alpha)$. For values off the a grid, we have used smoothing spline interpolation. The figure also displays a pointwise 95% confidence interval for the Bayes factor, obtained by bootstrapping the importance samples. In this plot, the minimum of $\log_{10}\{B(x_{1:n})\}$ over α is approximately -5 , i.e., $\min_{\alpha} B(x_{1:n}) \approx 10^{-5}$, showing negligible evidence in favor of the parametric null. This value is comparable to, but a bit smaller than, the minimum Bayes factor $\approx 7 \times 10^{-4}$ obtained in Berger and Guglielmi (2001) for their Polya tree alternative.

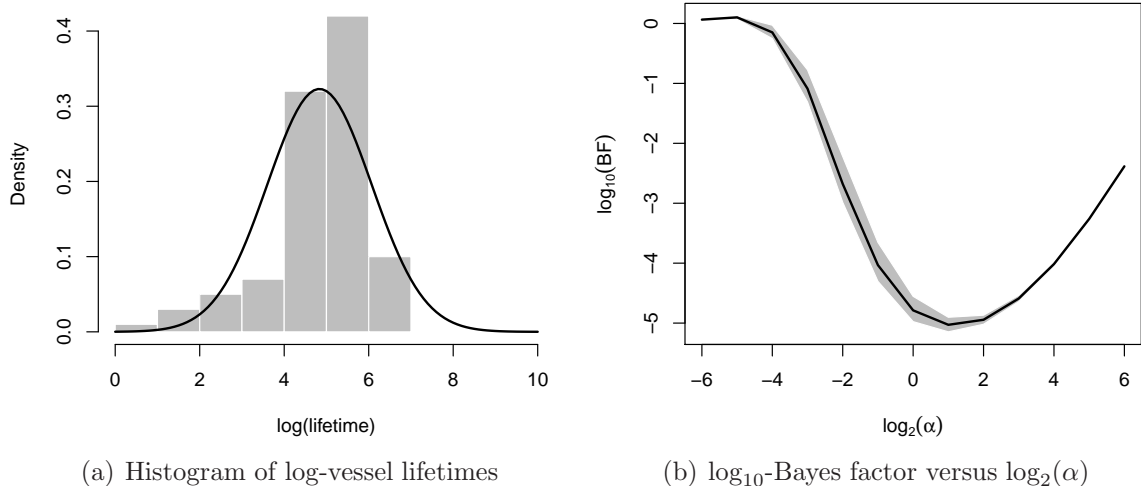


Figure 1: Results from Example 1. (a) Histogram with the best fitting normal density overlaid; (b) Bayes factor as a function of the Dirichlet process precision parameter α . The gray band represents a pointwise 95% bootstrap confidence interval.

The advantage of the Dirichlet process mixture alternative is that samples from this posterior have infinitely smooth densities with probability one. This is in contrast to the almost sure nowhere differentiability of samples from the Polya tree alternative in Berger and Guglielmi (2001). The next example illustrates that the test based on the Dirichlet process mixture alternative has greater discriminatory power than that based on the Polya tree alternative for models which are (mostly) smooth but non-normal.

Example 2. Consider three non-normal location–scale family models, namely, a Student-t distribution, a skew-normal distribution, and a uniform distribution. The first has heavier-than-normal tails, characterized by the degrees of freedom ν , the second has a skew, characterized by a shape parameter κ , and the third has discontinuities and no tails. Here we compare the discriminatory power of the proposed Dirichlet process mixture-based test to that of Berger and Guglielmi’s Polya tree-based test.

Computation of the Dirichlet process mixture-based Bayes factors, for this $p = 1$ case, is carried out as described Section 3 above, for α satisfying $\log_2(\alpha) \in -6 : 4$. For the Polya tree-based test, we implement the fixed-partition (Type 2) version as advocated in Berger and Guglielmi (2001, Equation 2) with the function $d(\varepsilon_m) = h^{-1}4^m$ for scale parameter h satisfying $\log_2(h) \in -6 : 4$. The Bayes factors reported below are the minima over the range of α and h , respectively.

Table 1 shows the Polya tree and Dirichlet process mixture Bayes factors for ten independent, randomly chosen data sets of size $n \in \{50, 100, 200\}$ from each of the $t(\nu = 3)$, $SN(\kappa = 10)$, and $Unif(-0.5, 0.5)$ models. Here we observe that the Bayes factors for the Dirichlet process mixture model are generally smaller than those for the Polya tree model—sometimes orders of magnitude smaller—which illustrates the former’s ability to better discriminate a non-normal true distribution from normal. For the Student-t and skew-normal, the Dirichlet process mixture test tends to produce smaller Bayes factors, so the conclusions one reaches, based on Jeffreys’ interpretation (Jeffreys 1961), are mostly

Model	$n = 50$		$n = 100$		$n = 200$	
	PT	DPM	PT	DPM	PT	DPM
t	0.94	0.91	1.01×10^0	9.20×10^{-1}	6.25×10^{-4}	4.72×10^{-3}
	0.60	0.25	6.91×10^{-4}	4.31×10^{-7}	5.38×10^{-2}	5.13×10^{-3}
	0.27	0.89	1.53×10^{-1}	3.96×10^{-4}	1.24×10^{-3}	4.28×10^{-4}
	0.002	0.0003	3.30×10^{-6}	4.30×10^{-9}	3.22×10^{-3}	7.43×10^{-4}
	0.98	0.59	3.30×10^{-3}	9.44×10^{-4}	1.44×10^{-8}	3.76×10^{-13}
	0.13	0.03	1.90×10^{-1}	1.44×10^{-4}	8.53×10^{-5}	1.27×10^{-6}
	1.01	0.96	1.03×10^0	6.40×10^{-1}	2.96×10^{-2}	5.07×10^{-6}
	0.54	0.05	1.02×10^{-6}	1.93×10^{-9}	2.56×10^{-13}	9.32×10^{-18}
	1.02	0.27	4.44×10^{-1}	4.77×10^{-2}	5.93×10^{-5}	8.29×10^{-8}
	0.004	0.001	3.86×10^{-4}	6.88×10^{-7}	1.63×10^{-5}	8.00×10^{-8}
SN	0.94	0.83	8.25×10^{-4}	1.32×10^{-4}	3.79×10^{-5}	3.60×10^{-6}
	1.01	0.06	6.38×10^{-1}	8.96×10^{-3}	3.84×10^{-7}	8.62×10^{-11}
	0.46	0.11	7.68×10^{-3}	7.04×10^{-4}	6.49×10^{-3}	1.02×10^{-5}
	0.11	0.08	4.40×10^{-2}	5.68×10^{-3}	3.25×10^{-2}	2.10×10^{-5}
	1.00	0.46	1.27×10^{-2}	3.77×10^{-4}	5.20×10^{-3}	7.59×10^{-5}
	1.01	0.34	2.38×10^{-1}	3.16×10^{-3}	3.50×10^{-3}	6.86×10^{-7}
	1.02	0.45	6.24×10^{-1}	3.48×10^{-3}	4.23×10^{-10}	1.44×10^{-13}
	0.38	0.11	7.65×10^{-7}	7.78×10^{-9}	3.14×10^{-4}	1.22×10^{-6}
	0.67	0.29	2.53×10^{-2}	1.07×10^{-4}	2.30×10^{-2}	2.17×10^{-5}
	0.20	0.25	4.91×10^{-4}	3.71×10^{-6}	8.83×10^{-7}	3.40×10^{-10}
Unif	0.29	0.01	0.04	1.40×10^{-7}	5.23×10^{-1}	4.53×10^{-6}
	0.20	0.04	0.51	3.88×10^{-3}	1.17×10^{-3}	4.63×10^{-9}
	1.02	0.14	0.25	1.22×10^{-2}	1.23×10^{-8}	1.68×10^{-13}
	1.01	0.01	1.03	1.17×10^{-5}	4.40×10^{-2}	5.03×10^{-7}
	0.62	0.01	1.02	1.02×10^{-2}	3.17×10^{-2}	2.25×10^{-6}
	0.83	0.04	0.04	1.28×10^{-3}	2.58×10^{-2}	3.56×10^{-8}
	0.93	0.18	0.17	7.93×10^{-6}	1.35×10^{-5}	2.89×10^{-8}
	0.96	0.38	0.76	1.82×10^{-2}	2.46×10^{-1}	5.36×10^{-5}
	1.01	0.04	0.09	6.92×10^{-3}	5.61×10^{-2}	3.94×10^{-5}
	1.01	0.09	1.04	1.02×10^{-1}	3.62×10^{-2}	1.37×10^{-5}

Table 1: Minimum Bayes factors for testing normality against Polya tree and Dirichlet process mixture alternatives, respectively. For each model and sample size configuration, as described in Example 2, Bayes factors are shown for ten random data sets.

the same for the two methods for $n = 50$ and $n = 200$. However, there are some striking differences in the $n = 100$ case, in particular, those two highlighted in bold. Histograms of the two data sets in question are shown in Figure 2, along with the best-fit normal density function overlaid. In both cases, normality is doubtful: the Student-t data in panel (a) has some extreme outliers relative to the corresponding normal, and the skew-normal data in panel (b) is highly asymmetric. Therefore, the strong evidence against the null hypothesis, as indicated by our proposed test, seems more reasonable.

The results in Table 1 for the uniform distribution are also quite surprising. Seemingly, the Polya-tree test should be better at detecting the discontinuities of the uniform distribution, but our results that show the Dirichlet process mixture test to be much more powerful are contrary to this intuition. One possible explanation is that a relative small sample from a uniform distribution can be better modeled by a mixture of narrow, densely packed smooth Gaussian densities than by some other nowhere smooth density.

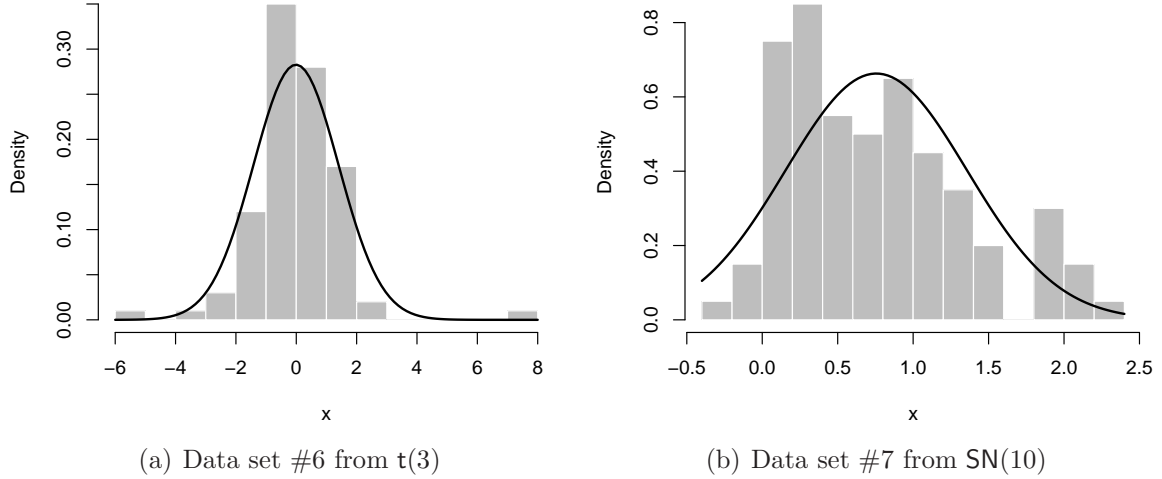


Figure 2: Histograms for the data sets corresponding to the two highlighted sets of Bayes factors in Table 1, along with the best-fit normal density function overlaid.

In Examples 1 and 2, we see that the proposed testing procedure tends to give little support to the null hypothesis of normality. As a last example in this univariate case, we shall investigate whether the proposed testing procedure tends to over-fit the data, favoring the alternative even when the null is true.

Example 3. Like in Example 2, we sample ten data sets of size $n \in \{50, 100, 200\}$. This time, though, the data sets are sampled from $N(0, 1)$, i.e., the null hypothesis is true. The minimum Bayes factors for the Polya tree and Dirichlet process mixture alternatives are shown in Table 2 over the same range of h and α as in Example 2. Here we find that the Polya tree test generally has a slightly larger Bayes factor than the Dirichlet process mixture test. This is not surprising, because a normal density is much closer to looking like a realization from a Dirichlet process mixture model than from a Polya tree model. But the fact that the two methods produce Bayes factors near unity indicates that neither is drastically over-fitting the data.

4.2 Multivariate case, $p > 1$

Example 4. Here we consider three different simulated data sets of size $n = 100$ in the $p = 2$ case. The first two models in question are a bivariate normal and a bivariate Student-t with three degrees of freedom. The third model, defined through a copula, has normal marginals but a non-elliptical joint distribution. Figure 3 shows scatterplots of the three data sets, all on the same scale, along with plots of the proposed Bayes factor versus the Dirichlet precision parameter α . In the normal case (top row) we see that the Bayes factor never drops below 1, so there would be no reason to doubt the normality assumption. The second row shows a Student-t data set where we see a high concentration of points around $(0, 0)$ along with a number of potential outliers, suggesting a heavier-than-normal tail. Since the Bayes factor for this case bottoms out just below 10^{-5} , our proposed testing procedure is apparently able to pick up the effect of the heavier

Model	$n = 50$		$n = 100$		$n = 200$	
	PT	DPM	PT	DPM	PT	DPM
$N(0, 1)$	0.42	0.97	1.03	0.99	1.06	0.91
	1.00	0.68	1.02	1.02	1.07	0.90
	0.86	0.75	1.03	1.00	1.05	0.95
	1.02	0.92	0.93	0.89	0.97	0.97
	1.01	0.97	1.02	0.91	1.07	0.90
	1.02	0.98	0.75	0.84	1.09	0.94
	1.02	0.95	1.05	0.94	1.09	0.95
	1.02	0.97	1.02	0.91	1.06	0.99
	1.03	0.99	0.65	0.91	0.92	0.61
	1.02	0.96	1.02	0.93	1.11	0.99

Table 2: Minimum Bayes factors for testing normality against Polya tree and Dirichlet process mixture alternatives, respectively. These are shown for ten random data sets from a standard normal distribution.

than expected tail. Finally, based on the sharp decrease in the Bayes factor in the last row, it is apparent that the proposed test can easily pick up the non-elliptic shape of the underlying distribution.

Example 5. As a last illustration, we investigate the development of ancient Egyptian skulls. Our data consists of $p = 4$ measurements—maximal breadth, basibregmatic height, basialveolar length, and nasal height—taken on $n = 150$ ancient skulls at one of five time periods ranging from 4000 B.C. to 200 A.D. These data appear in Thomson and Randall-Maciver (1905); see also <http://lib.stat.cmu.edu/DASL>. To test for differences in mean skull shape across the five time periods, it is standard to perform a multivariate analysis of variance, or MANOVA. The MANOVA results (not shown) indicate that there is a significant difference between the mean skull measurements across time; however, the validity of this procedure assumes the error terms to have a multivariate normal distribution. Here we apply the proposed Bayesian methodology to the residuals to assess the assumption of normality. Figure 4(a) shows a quantile plot of the observed Mahalanobis distance scores versus the theoretical chi-square null distribution. A visual inspection of this plot suggests that there may be some doubt about the postulated chi-square model, though apparently not substantial. But this plot is a one-dimensional summary of a four-dimensional distribution, so some information may be lost. For a more complete summary, Figure 4(b) shows the estimated Bayes factor against the Dirichlet precision parameter α . The \log_{10} -Bayes factor bottoms out around -0.83 which suggests that there is only a small amount of evidence against the null hypothesis. However, due the magnitude of the Bayes factor for moderate to large α values, any reasonable weighted average over α would again produce a large Bayes factor. Therefore, we conclude that there is no significant evidence against the four-dimensional normality of the MANOVA residuals.

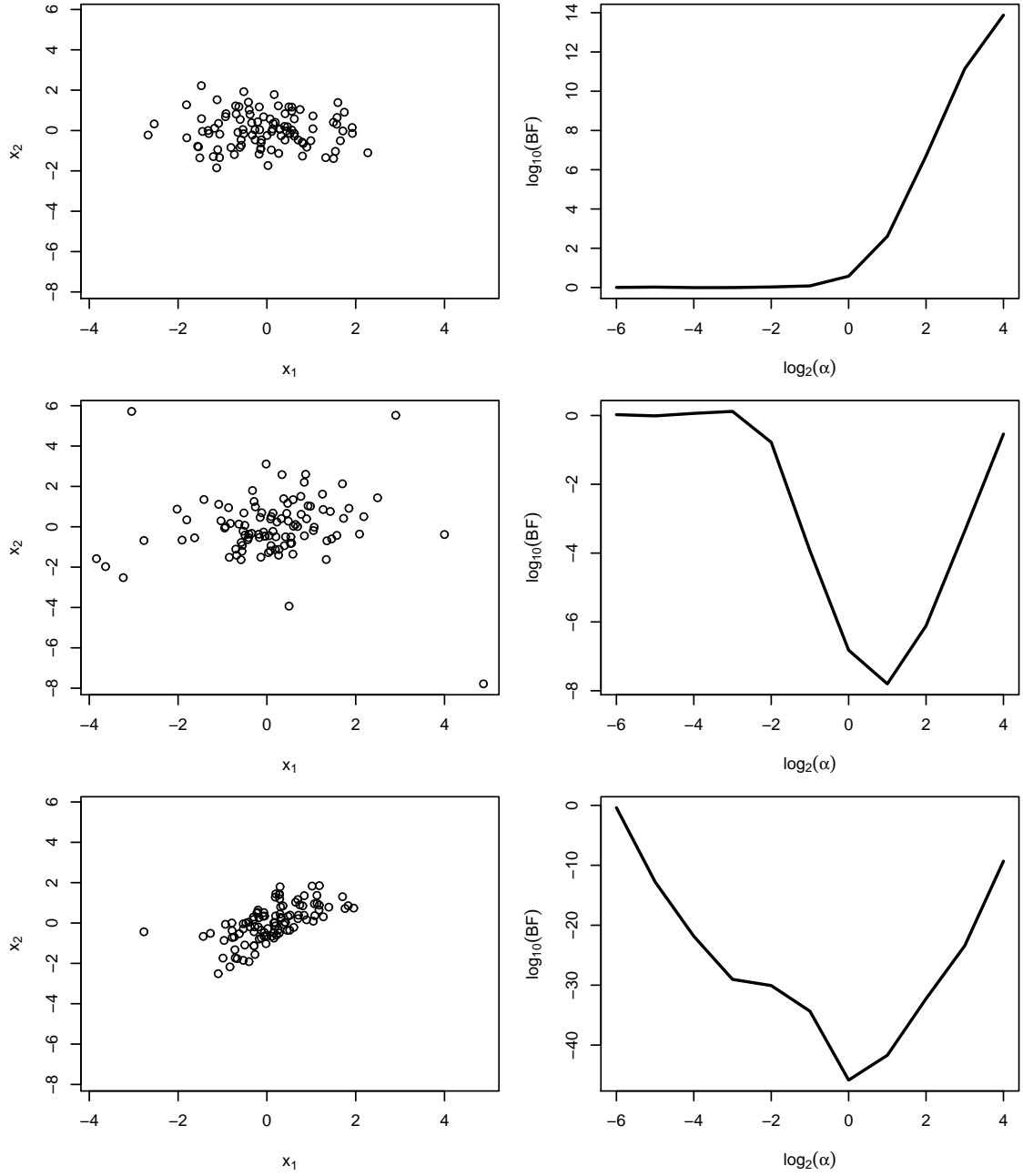


Figure 3: Data scatterplots and plots of the Bayes factor versus α as in Example 4. The first row is normal; the second row is Student-t; the third row is non-elliptical.

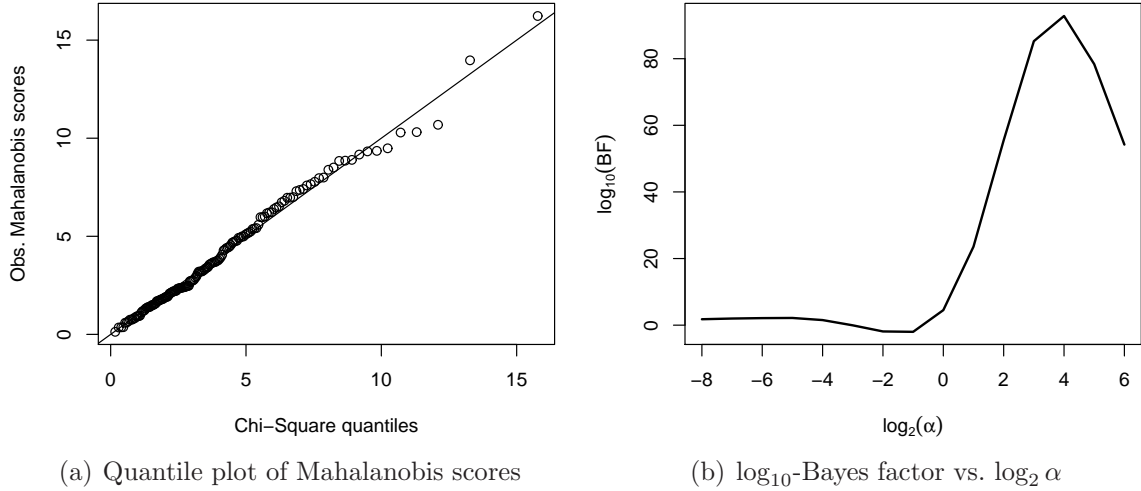


Figure 4: Two summaries of the evidence for/against the normality null hypothesis in the ancient Egyptian skull problem described in Example 5.

5 Concluding remarks

This paper presents a novel approach to testing p -variate normality, for general $p \geq 1$, from a Bayesian perspective. In particular, we propose a broad nonparametric alternative based on a Dirichlet process mixture prior, and show that the null normal model is suitably embedded in this class. A predictive matching property of this new alternative class is established, which justifies the use of a common, noninformative prior for the nuisance location and scale parameters in both the null and alternative hypotheses. Consequently, the proposed test requires the user to specify only a single scalar tuning parameter. An efficient sequential importance sampler is provided which allows for fast evaluation of Bayes factors over a range of tuning parameter values. Simulation results demonstrate that the proposed method has higher discriminatory power when the true, data-generating distribution is a smooth departure from normality, and also avoids over-fitting when the true distribution is normal.

An important and challenging open question is if large-sample consistency of the resulting Bayes factor obtains. For this, one first needs a posterior consistency result for the proposed nonparametric alternative model. The authors believe that an argument along the lines given in Wu and Ghosal (2010) can be used to establish posterior consistency, but this is only half the battle for Bayes factor consistency. As mentioned in Tokdar et al. (2010), because the nonparametric model can recover the a true normal model nearly as efficiently as a parametric model, special conditions are needed (see McVinish et al. 2009) to prove Bayes factor consistency under H_0 , and it is not yet clear if even standard priors satisfy these conditions.

Acknowledgment

A portion of this work was completed while R. Martin was with the Department of Mathematical Sciences, Indiana University–Purdue University Indianapolis.

References

- Andrews, D. F. and Herzberg, A. M. (1985), *Data*, New York: Springer-Verlag.
- Berger, J. O. and Guglielmi, A. (2001), “Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives,” *J. Amer. Statist. Assoc.*, 96, 174–184.
- Berger, J. O., Pericchi, L. R., and Varshavsky, J. A. (1998), “Bayes factors and marginal distributions in invariant situations,” *Sankhyā Ser. A*, 60, 307–321.
- Blackwell, D. and MacQueen, J. B. (1973), “Ferguson distributions via Pólya urn schemes,” *Ann. Statist.*, 1, 353–355.
- Carota, C. and Parmigiani, G. (1996), “On Bayes factors for nonparametric alternatives,” in *Bayesian statistics, 5 (Alicante, 1994)*, New York: Oxford Univ. Press, Oxford Sci. Publ., pp. 507–511.
- Castillo, I. (2008), “Lower bounds for posterior rates with Gaussian process priors,” *Electron. J. Stat.*, 2, 1281–1299.
- Choudhuri, N., Ghosal, S., and Roy, A. (2005), “Bayesian methods for function estimation,” in *Bayesian thinking: modeling and computation*, Elsevier/North-Holland, Amsterdam, vol. 25 of *Handbook of Statist.*, pp. 373–414.
- DasGupta, A. (2008), *Asymptotic Theory of Statistics and Probability*, New York: Springer.
- Dykstra, R. L. and Hewett, J. E. (1978), “Positive dependence of the roots of a Wishart matrix,” *Ann. Statist.*, 6, 235–238.
- Escobar, M. D. (1994), “Estimating normal means with a Dirichlet process prior,” *J. Amer. Statist. Assoc.*, 89, 268–277.
- Escobar, M. D. and West, M. (1995), “Bayesian density estimation and inference using mixtures,” *J. Amer. Statist. Assoc.*, 90, 577–588.
- Ferguson, T. S. (1973), “A Bayesian analysis of some nonparametric problems,” *Ann. Statist.*, 1, 209–230.
- Florens, J.-P., Richard, J.-F., and Rolin, J.-M. (1996), “Bayesian encompassing specification tests of a parametric model against a nonparametric alternative,” Tech. Rep. 96.08, Université Catholique de Louvain, Institut de Statistique.

- Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999), “Posterior consistency of Dirichlet mixtures in density estimation,” *Ann. Statist.*, 27, 143–158.
- Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000), “Convergence rates of posterior distributions,” *Ann. Statist.*, 28, 500–531.
- Ghosal, S. and van der Vaart, A. W. (2001), “Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities,” *Ann. Statist.*, 29, 1233–1263.
- (2007), “Posterior convergence rates of Dirichlet mixtures at smooth densities,” *Ann. Statist.*, 35, 697–723.
- Ghosh, J. K., Delampady, M., and Samanta, T. (2006), *An introduction to Bayesian analysis*, New York: Springer.
- Ghosh, J. K. and Ramamoorthi, R. V. (2003), *Bayesian Nonparametrics*, New York: Springer-Verlag.
- Jeffreys, H. (1961), *Theory of probability*, Third edition, Clarendon Press, Oxford.
- Kass, R. E. and Raftery, A. E. (1995), “Bayes factors,” *J. Amer. Statist. Assoc.*, 90, 773–795.
- Lavine, M. (1992), “Some aspects of Pólya tree distributions for statistical modelling,” *Ann. Statist.*, 20, 1222–1235.
- (1994), “More aspects of Pólya tree distributions for statistical modelling,” *Ann. Statist.*, 22, 1161–1176.
- Liu, J. S. (1996), “Nonparametric hierarchical Bayes via sequential imputations,” *Ann. Statist.*, 24, 911–930.
- (2001), *Monte Carlo strategies in scientific computing*, New York: Springer-Verlag.
- Lo, A. Y. (1984), “On a class of Bayesian nonparametric estimates. I. Density estimates,” *Ann. Statist.*, 12, 351–357.
- MacEachern, S. and Müller, P. (1998), “Estimating mixture of Dirichlet process models,” *J. Comput. Graph. Statist.*, 7, 223–238.
- MacEachern, S. N. (1998), “Computational methods for mixture of Dirichlet process models,” in *Practical nonparametric and semiparametric Bayesian statistics*, New York: Springer, pp. 23–43.
- Mauldin, R. D., Sudderth, W. D., and Williams, S. C. (1992), “Pólya trees and random distributions,” *Ann. Statist.*, 20, 1203–1221.
- McVinish, R., Rousseau, J., and Mengersen, K. (2009), “Bayesian goodness of fit testing with mixtures of triangular distributions,” *Scand. J. Stat.*, 36, 337–354.

- Muirhead, R. J. (1982), *Aspects of multivariate statistical theory*, New York: John Wiley & Sons Inc., Wiley Series in Probability and Mathematical Statistics.
- Müller, P. and Quintana, F. A. (2004), “Nonparametric Bayesian data analysis,” *Statist. Sci.*, 19, 95–110.
- Neal, R. M. (2000), “Markov chain sampling methods for Dirichlet process mixture models,” *J. Comput. Graph. Statist.*, 9, 249–265.
- Sethuraman, J. (1994), “A constructive definition of Dirichlet priors,” *Statist. Sinica*, 4, 639–650.
- Thomson, A. and Randall-Maciver, R. (1905), *Ancient Races of the Thebaid*, Oxford: Oxford University Press.
- Tokdar, S. T., Chakrabarti, A., and Ghosh, J. K. (2010), “Bayesian nonparametric goodness of fit tests,” in *Frontiers of Statistical Decision Making and Bayesian Analysis*, eds. Cheh, M.-H., Dey, D., Müller, P., Sun, D., and Ye, K., Springer, pp. 185–194.
- van der Vaart, A. W. and van Zanten, J. H. (2008), “Rates of contraction of posterior distributions based on Gaussian process priors,” *Ann. Statist.*, 36, 1435–1463.
- Verdinelli, I. and Wasserman, L. (1998), “Bayesian goodness-of-fit testing using infinite-dimensional exponential families,” *Ann. Statist.*, 26, 1215–1241.
- Wu, Y. and Ghosal, S. (2010), “The L_1 -consistency of Dirichlet mixtures in multivariate Bayesian density estimation,” *J. Multivariate Anal.*, 101, 2411–2419.